© Health Research and Educational Trust
DOI: 10.1111/j.1475-6773.2006.00551.x

Use of Geocoding and Surname Analysis to Estimate Race and Ethnicity

Kevin Fiscella and Allen M. Fremont

Objective. To review two indirect methods, geocoding and surname analysis, for estimating race/ethnicity as a means for health plans to assess disparities in care.
Study Design. Review of published articles and unpublished data on the use of geocoding and surname analyses.
Principal Findings. Few published studies have evaluated use of geocoding to estimate racial and ethnic characteristics of a patient population or to assess disparities in health care. Three of four studies showed similar estimates of the proportion of blacks and one showed nearly identical estimates of racial disparities, regardless of whether indirect or more direct measures (e.g., death certificate or CMS data) were used. However, accuracy depended on racial segregation levels in the population and region assessed and geocoding was unreliable for identifying Hispanics and Asians/Pacific Islanders. Similarly, several studies suggest surname analyses produces reasonable estimates of whether an enrollee is Hispanic or Asian/Pacific Islander and can identify disparities in care. However, accuracy depends on the concentrations of Asians or Hispanics in areas assessed. It is less accurate for women and more acculturated and higher SES persons due intermarriage, name changes, and adoption. Surname analysis is not accurate for identifying African Americans. Recent unpublished analyses suggest plans can successfully use a combined geocoding/surname analyses approach to identify disparities in care in most regions. Refinements based on Bayesian methods may make geocoding/surname analyses appropriate for use in areas where the accuracy is currently poor, but validation of these preliminary results is needed.
Conclusions. Geocoding and surname analysis show promise for estimating racial/ethnic health plan composition of enrollees when direct data on major racial and ethnic groups are lacking. These data can be used to assess disparities in care, pending availability of self-reported race/ethnicity data.

Key Words. Quality indicators, health care, managed care programs, data collection, continental ancestry, outcome and process assessment health care

Eliminating racial and ethnic disparities in health care will require health care plans, hospitals, and other health care organizations to obtain race and ethnicity on their plan members or patients (Institute of Medicine 2002; National

Quality Forum 2002; Workgroup on Quality: National Committee on Vital and Health Statistics 2004). Obtaining such data on all members represent a formidable task, particularly for health plans, who generally have not retained racial/ethnic data on enrollees because of uncertainty about the legality of collecting racial/ethnic data or fear that consumers would assume plans were misusing the data (Fremont and Lurie 2004).

Clarifications about the legality of collecting race/ethnicity data (Perot and Youdelman 2001; Rosenbaum and Painter 2005) and positive reactions to plans' efforts to collect race/ethnicity information from enrollees to assess disparities in care have helped allay plans' concerns and increase their interest in obtaining race/ethnicity (Hassett 2005; Nerenz 2005). Nonetheless, plans have limited options for quickly obtaining needed data. Direct (e.g., self-report) race/ethnicity data could be solicited at the time of enrollment for *new* members, but alternative strategies are needed to obtain this information from *existing* plan members. Potential methods include mail, telephone, or Internet surveys; onsite collection at point of care; and supply by employers, hospitals, states, or Centers for Medicare and Medicaid Services. However, each of these strategies has limitations in terms of reliability, validity, bias, and completeness. In addition, obtaining race/ethnicity data using these direct methods typically takes plans a number of years to complete for most of their enrollees.

As a practical matter, no method for obtaining race and ethnicity data can be entirely accurate or bias free. Surveys are limited by nonresponse. For example, Aetna reports that 20 percent of its plan members refuse to provide these data voluntarily (Jack Rowe, personal communication). Whether these refusals differ by race and ethnicity is unknown, but indirect methods could be used to assess the potential bias here. More generally, response rates to all types of surveys have consistently declined (Bickart and Schmittlein 1999). Response rate to the Consumer Assessment of Health Plan Survey® (CAPHS) is often under 50 percent (Weech-Maldonado et al. 2003) and differs by race and education (Zaslavsky, Zaborski, and Cleary 2002).

Furthermore, hospitals' inconsistent collection and classification of race and ethnicity data is potentially problematic. For example, one study reports that race was coded differently upon rehospitalization 6 percent of rehospitalized African Americans and 11 percent of rehospitalized whites (Blustein

---

Address correspondence to Kevin Fiscella, M.D., MPH, Associate Professor, Departments of Family Medicine and Community & Preventive Medicine, University of Rochester, School of Medicine & Dentistry, Rochester, NY 14620. Allen M. Fremont, M.D., Ph.D., Natural Scientist and Sociologist, is with the RAND Corporation, Santa Monica, CA.

1994). Another study found that race and ethnicity compiled by the Veterans Administration Health System corresponds with self-report data only 60 percent of the time, with lower agreement for nonwhites and better educated patients (Kressin et al. 2003).

In this paper, we provide a guide for health plans and other health care organizations who want to begin assessing racial/ethnic disparities among their members, but lack racial/ethnic data on most or all of their enrollees. In particular, we discuss two indirect methods for estimating race/ethnicity—geocoding and surname analyses—that potentially provide plans with an inexpensive and efficient means to estimate racial/ethnic composition of their members, and identify and target disparities while they proceed with the longer process of collecting direct race/ethnicity data.

Geocoding refers to the use of plan members' home address to infer other information about them, including their race and ethnicity. Surname analysis refers to the use of last names for similar purposes. We address the advantages, accuracy, and limitations of these methods and offer practical suggestions for using them.

GEOCODING

Geocoding involves using plan members' addresses to identify geographic areas where they live and linking this information to U.S. Census Bureau data about that area. Census data includes rates of poverty, educational levels, employment levels, and racial/ethnic composition of areas. Thus, geocoded measures can be used, with varying levels of accuracy, to infer characteristics about persons living in those areas such as their likely race, in addition to area socioeconomic status (SES). Linkage to other data sources can provide additional information regarding the physical and environmental properties of neighborhoods such as levels of environmental lead, crime, traffic patterns, walking or bike paths, and liquor stores.

Geocoding can be performed at different geographic levels, but degree of correspondence between area and individual characteristics generally increases when smaller, more homogenous units of analysis are used (Krieger et al. 2002). Zip codes generally span relatively large geographic areas, often including 30,000 or more people from widely varying sociodemographic subgroups. Census tracts are smaller, averaging about 4,000 residents, and tend to be more homogenous although it is not uncommon for the same tract to include both affluent and poor neighborhoods. Census block-groups

average 1,000 or fewer residents—the size of a small neighborhood—and generally are quite homogenous. Finally, census blocks are small areas demarcated by visible boundaries such as streets, streams, and railroad tracks and also quite homogenous, although the number of residents residing there can be very small—averaging about 50 residents in the two-thirds of the blocks that are populated (Krieger, Williams, and Moss 1997; U.S. Census Bureau 2004).

Uses of Geocoding in Health Care

Geocoding has been used for decades, but its use in monitoring quality of care, is relatively new. Researchers routinely use geocoding to identify sociodemographic characteristics of populations and areas as a means of targeting vulnerable populations or to estimate effects of sociodemographic characteristics on health when individual-level data are unavailable (Krieger, Williams, and Moss 1997; Geronimus and Bound 1998; Elreedy et al. 1999). Geocoded measures are commonly used to estimate SES including percent poverty or median income when direct data are lacking (Krieger 1992). Geocoding is sometimes used to determine latitude and longitude of a patient's residence and, in conjunction with geographic information systems, to calculate distance or time to travel to providers—both of which are potential barriers to care (Institute of Medicine 2002). In addition, geocoding can be used to assess neighborhood characteristics, such as the availability of a public transportations system, grocery stores, parks, and crime rates that are associated with health behaviors and health outcomes independent of individual characteristics (Elreedy et al. 1999; Heck, Schoendorf, and Chavez 2002).

Marketing departments within some health care organizations and commercial vendors have become adept at using geocoding (and surname analyses described below) to identify populations of interest or profile sociodemographic characteristics of potential service areas. However, because of concerns about potential misuse of data to preferentially enroll or differentially treat certain racial, ethnic, or SES groups, this type of information typically has not been shared with clinical or quality management staff.

Accuracy and Validity

Most validation work on geocoded measures has focused on geocoded measures of SES, sometimes referred to as area-based socioeconomic measures (Krieger et al. 2003b). These estimates generally correspond well to estimates based on self-reported SES and show similar associations with different health

outcomes (Fiscella and Franks 2001; Krieger et al. 2002; Krieger, Chen et al. 2003; Krieger, Waterman et al. 2003). Census block-group and tract measures performed comparably in detecting SES gradients. However, zip codes failed to detect effects or showed opposite effects on numerous health indicators in some (Krieger, Waterman et al. 2003), but not all studies (Fiscella and Franks 2001).

We found only a handful of published studies that assessed the validity of geocoding to estimate race and ethnicity or examine the association of these estimates with health or quality of care indicators (Andjelkovich et al. 1990; Kwok and Yankaskas 2001; Chen et al. 2002; Fremont and Wickstrom 2002). Although the results of these studies are mixed, some conclusions seem clear.

First, the accuracy of geocoded estimates of race and ethnicity largely depends on the extent of racial and ethnic segregation in the geographic areas considered. Greater proportions of particular minority groups living in racially *segregated* areas yield higher specificity (i.e., lower false-positive rate) of geocoded estimates. Conversely, higher proportions of minorities living in *integrated* neighborhoods yield lower sensitivity (i.e., higher false-negative rate). Furthermore, accuracy of geocoded estimates vary somewhat depending on the geographic level (i.e., Census tract, block group, or block) used. Theoretically, accuracy should be greatest with smallest areas (e.g., blocks), but limited data suggest that *aggregated* estimates of racial/ethnic composition and disparities are similar regardless of geographic unit. When plans serve significant number of enrollees in less dense or rural areas, we recommend performing geocoding at the Census block group or tract level rather than at the block level because the number of residents in many blocks may be too small to generate reliable estimates.

Because at least half of black Americans continue to live in predominately black neighborhoods (Glaeser and Vigdor 2001), geocoding can produce reasonably accurate estimates of black race in many areas of the country. Fremont, Bierman et al. (2005) used geocoding to identify likely black Medicare+Choice enrollees from four regions based on whether they lived in block groups where more than two-thirds of the residents were black. Among 17,500 enrollees, 92 percent were successfully matched to block groups and "identified" as either black or other (other group was mainly white). Race data from CMS files were used to determine "true" race. Of those assigned race based on geocoding, 89 percent were classified correctly. Most (86 percent) of the 11 percent of enrollees incorrectly identified as nonblack, did not live in predominantly black neighborhoods. Similarly in a Detroit study, investigators correctly classified whether individuals were black or white for

91 percent of a sample ($N = 310$) geocoded to the block group (Andjelkovich et al. 1990).

Geocoded measures of race are less reliable in more integrated regions. Chen, Petitti, and Enger (2004) found that only 47,328 (30.6 percent) of 117,209 Southern California Kaiser members identified as black on hospital discharge files lived in a Census block with more than 50 percent black residents (Chen, Petitti, and Enger 2004). Similarly, geocoded measures did not reliably predict which women in a North Carolina mammography registry were black (Kwok and Yankaskas 2001). In general, geocoded race measures perform much better in highly segregated cities such as Detroit, than less segregated cities such as San Diego or Raleigh–Durham–Chapel-Hill (Robinson and Grant-Thomas 2004). Because Hispanics, Asians, and many Native Americans tend to live in far less segregated neighborhoods than blacks (Massey and Denton 1989; Logan 2001), geocoding, alone, generally is not useful for identifying members of these minority groups.

The accuracy of geocoding for "whites" has not been systematically examined in published studies. More typically, blacks are compared with "nonblack." Preliminary unpublished results suggest that geocoding alone overestimates white numbers by several percentage points, but yields comparable composition and disparities estimates compared with more direct measures (Fremont, Pantajo et al. 2005).

The accuracy of geocoded measures of race/ethnicity also depends on the "cut point" (e.g., percent of residents black) used to classify individuals as black. Increasing the cut point decreases the false positive rate, but increases the false-negative rate. The optimal cut point depends on the distribution of racial/ethnic groups in the areas considered and the purpose for the estimates, but typically ranges between 50 and 75 percent.

The validity of geocoding to infer race depends on its purpose. It is not sufficiently accurate to infer *individual* race. However, it seems to be valid for estimating racial/ethnic composition of plans and assessing indicators of care outcomes for groups *in the aggregate.* Chen et al. (2002) found that even in relatively racially integrated areas like Southern California, geocoded estimates yielded reasonable estimates of the proportion of black Kaiser members in the region overall and within different Medical Centers (Table 1). Fremont, Bierman et al. (2005) showed that geocoded measures for blacks yielded nearly identical estimates of disparities in quality of care as those using race derived from Medicare data (Table 2). Fremont, Bierman et al. have also used geocoded measures of race among commercial plan enrollees to demonstrate racial disparities in HEDIS performance (Fremont and Wickstrom 2002) and

Table 1: Percentage of Blacks among Women Giving Birth, from Birth Certificates and as Implied by Geocoded Data, for Selected Kaiser Medical Centers in Southern California (S. CA)*

Medical Center	*N*	*Birth Certificate (%)*	*Geocoded (to Block Level) (%)*
All S. CA centers	141,537	46	42
Selected S. CA medical centers			
San Diego	22,877	7	6
Woodland Hills	5,633	4	3
Bellflower	14,804	13	9
West LA	8,937	48	40
Baldwin park	10,178	4	3

*Adapted from Chen et al. (2004).

in use of cardiovascular diagnostic tests and therapeutic procedures (Fremont, Wickstrom, and Escarce 2003). Geocoding is not sufficiently accurate to send out individual letters to plan members implying that they have been identified as black. Finally, estimates of racial disparities in health may differ between those based on geocoded and individual measures not only because of racial misclassification, but also because community-level factors (e.g., safe areas to exercise or lead exposure), may differ depending on area racial composition.

SURNAME ANALYSIS

Surname analysis uses an individual's last name to estimate the likelihood that the individual belongs to a particular racial or ethnic group. Surname analysis is more reliable for identifying Hispanics and Asians than African Americans because of more distinctive last names among the former groups.

The U.S. Census Bureau has used Spanish surnames to the identify Hispanics for nearly 50 years (Word and Perkins 1996). Surname analysis has been used to assess mortality (Rosenwaike, Hempstead, and Rogers 1991), cancer incidence (Swallen et al. 1997, 1998; Coronado et al. 2002), rates of cancer screening among HMO enrollees (Jacobs and Lauderdale 2001), local concentrations of ethnic groups (Rosenwaike 1994), the ethnic composition of homeowners, and the ethnicity of patients (Coronado et al. 2002; NAACCR Expert Panel on Hispanic Identification 2003). Marketing and political consulting companies use variations of this technique to identify race/ethnicity of

Table 2: Comparison of Racial Disparities on Process Measures for Medicare+Choice Plan Enrollees Using CMS-Based and Geocoded Race/Ethnicity Information

	Enrollees Classified by CMS as (%)			*Enrollees Classified by Geocoding (%)*		
Measures	*Black*	*Nonblack*	*Difference*	*Mostly Black Neighborhoods*	*Mostly Nonblack Neighborhoods*	*Difference*
Beta-blocker after myocardial infarction	37.8	60.6	−22.7[a]	40.0	63.1	−23.1[a]
LDL after cardiac event	45.2	62.4	−17.2[b]	45.3	61.9	−16.5[a]
HgbA1c check for diabetics	51.1	65.6	−14.5[b]	49.9	64.4	−14.5[b]
LDL check for diabetics	39.3	60.9	−21.6[b]	37.8	59.1	−21.3[b]
Eye exam for diabetics	39.0	46.6	−7.6[b]	38.6	45.9	−7.3[b]
Urine protein check for diabetics	19.8	18.0	1.9[a]	19.4	18.3	1.1

Source: Person-level race data from the Center for Medicare and Medicaid Services (CMS); Geocoded race data are from Managed Care Organization's administrative database; U.S. Census Bureau Data.

Notes: "Predominantly black" defined as neighborhoods in which >67 percent or more of residents were black. We coded race/ethnicity information from CMS as either black or some other race (i.e., nonblack) for consistency with the geocoded measures. In addition, coding CMS race/ethnicity data in this way is reasonable because although the black category is accurate in CMS data, but the white category includes a substantial number of nonblack Hispanics and other minorities.

[a] $p < .01$;

[b] $p < .001$.

potential consumers or voters (Abrahamse, Morrison, and Bolton 1994; Lee and Sutton 2002).

Types of Surname Analysis

There are several ways to assign ethnicity based on names: use of letter combinations, dictionaries of surnames, and combinations of first, middle, and last names. The original approach, the Generally Useful Ethnicity Search System (GUESS), was developed using 1953 California Department of Public Health birth data (Perez-Stable et al. 1995). The program was derived using an algorithm based on common Spanish names, given name, and mother's maiden name. It uses the linguistic structure of the last name to assign Hispanic ethnicity. GUESS was updated in the 1980s using more current Spanish surnames

(Rosenwaike and Bradshaw 1988). A simpler approach is to assign Hispanic ethnicity using a surname list. Such a list was developed using 1980 Census data (Perkins 1993) and then revised using 1990 data (Word and Perkins 1996).

Surname lists have also been used to identify Asian subpopulations in the United Kingdom (U.K.) (Nicoll, Bassett, and Ulijaszek 1986; Harland et al. 1997; Nanchahal et al. 2001), Australia (Hage et al. 1990), Canada (Coldman, Braun, and Gallagher 1988; Choi et al. 1993; Sheth et al. 1997), and the U.S. (Swallen et al. 1998; Lauderdale and Kestenbaum 2000). The best validated list has been produced by Lauderdale and Kestenbaum (2000) using the Social Security Administration's file. Separate surname lists have been generated for Chinese, Indian, Japanese, Korean, Filipino, and Vietnamese Americans.

Software has been developed to identify Asian subgroups using names, some of which incorporate first names. Nanchahal et al. (2001) developed a computer algorithm, the South Asian Names and Group Recognition Algorithm (SANGRA), that generates four outputs including South Asian ethnicity; religious affiliation; likely language; and whether ethnicity was assigned on the basis of surname and first name, surname only, first name only, or middle name only. Another system named "Nam Pehchan" relies on both stems as well as full matching of names (Martineau and White 1998; Cummins et al. 1999; Harding, Dews, and Simpson 1999). Although both systems performed reasonably well among U.K. samples, performance in the U.S. is unknown.

First names have been incorporated along with surnames in several national studies of Hispanic (Elo et al. 2004) and Asian (Lauderdale and Kestenbaum 2002) mortality. Commercial vendors have developed complex algorithms that incorporate middle names in addition to first and last names into coding schemes; the incremental benefit of this approach is not known.

Accuracy and Validity

Assessment of Hispanic and Asian ethnicity based on surname analysis has been shown to be reasonably accurate across diverse populations that contain adequate numbers of the ethnic group being assessed. Accuracy can be assessed based on sensitivity, specificity, and positive predictive value (percentage of persons with a given surname who self-report the ethnicity assigned by the coding method). The 1990 Census Spanish list (containing fewer than 1,000 Spanish surnames) showed an overall sensitivity of 79 percent and a specificity of 90 percent compared with self-reported ethnicity in a national sample (Perkins 1993).

Asian surnames yield similar overall accuracy. Lauderdale and Kestenbaum's name list, derived from Social Security records, and validated using the 1990 Census, showed sensitivities ranging from 74 percent for Vietnamese to 29 percent for Filipinos and positive predictive values ranging from 92 percent for Japanese to 76 percent for Chinese. Accuracy improves when race data are also available, but Asian ethnicity data are lacking. For example, availability of Asian race data can be used to distinguish Filipinos from Hispanics (Lauderdale and Kestenbaum 2000). When race was available, the sensitivity and positive predictive values for Filipinos increased from 29 to71 percent and 86 to 93 percent, respectively.

Names such as "Lee" or "Real" are less distinctive. Errors also occur because of intermarriage, name change, and adoption. Women who marry outside their ethnic group may be miscoded (Winkleby and Rockhill 1992; Perkins 1993). Rates of intermarriage differ by ethnicity, place of birth, acculturation, and SES (Winkleby and Rockhill 1992; Lauderdale and Kestenbaum 2000). For this reason, the sensitivity and specificity of surnames are higher for men and youth and lower SES persons. Spanish surnames have a sensitivity and specificity of 82 and 92 percent for men versus 77 and 88 percent for women (Perkins 1993). The prevalence of members of a particular ethnic group in the community has a powerful effect on surname accuracy. Sensitivity and specificity for Spanish surnames range from 88 and 96 percent in Texas to 34 and 37 percent in Vermont (Perkins 1993). No published data were found on use of surname analysis for identifying non-Hispanic whites, African Americans, or Native Americans.

Few data have been published using surname analysis to examine quality of care. Jacobs and Lauderdale (2001) reported no differences in cancer screening among HMO members between Hispanics and non-Hispanics using surname analysis. Nerenz et al. (2002) reported ethnic disparities in health care quality among health plan members in three of thirteen measures—disparities comparable to those reported in studies using self report measures.

Surname analysis, like geocoding, is potentially useful for assessing outcomes or measures related to racial/ethnic groups in the aggregate, but not for identifying the individual race/ethnicity of plan members. The sensitivity and positive predictive values of the algorithms and lists can often be adapted to the purpose at hand. If the aim is to capture as much of the population as possible such as over sampling for a survey, lower cut-points (inclusion of less distinctive names) should be used. If the aim is to ensure the highest positive predictive values, only names with very high probability of correct matching should be included.

COMBINED METHODS

The advantages and limitations of geocoding and surname analysis complement each other (Table 3) making combined use an attractive means for inferring race/ethnicity among plan members (Table 3). Geocoding is more reliable for inferring black race whereas surname analysis is better for inferring Hispanic or Asian ethnicity. Furthermore, geocoding provides estimates of the racial/ethnic composition of the area where surnames are applied.

When the two methods are applied to the same geographic area (e.g., census tract, block group, or block), overall accuracy can be improved. For example, a combined approach can improve the accuracy of geocoding of non-Hispanic African Americans and whites. Names assigned to Hispanic or Asian ethnicity can be removed from name lists used to assign white or black race, thereby refining such name lists for estimating the non-Hispanic white or black population. Incorrect assignment of minorities to the majority white population will have relatively little effect in most instances because of much higher numbers of white, non-Hispanics.

Conversely, the accuracy of geocoding can be improved by using information from geocoding (e.g., the percentage of each racial/ethnic group in a census tract, block group, or block). This information can be used to generate prior probabilities before assigning ethnicity based on surnames. Preliminary analyses using prior probabilities to refine estimates based on Bayes theorem suggest marked improvement in accuracy surname analysis alone or in combination with geocoding to identify likely blacks without the Bayes approach (Elliott et al. 2005). A combined approach can also help distinguish Hispanic subpopulations. For example, if the Hispanic population in a Census blockgroup is 90 percent Cuban, a person with a Spanish surname residing in that area can be coded as Cuban with high accuracy.

The combined approach also allows health plans to estimate the proportion of plan members who speak languages other than English. Surname analysis provides estimates of the number of plan members by ethnicity that live in a particular Census tract; Census data provide data on language spoken at home by persons of differing race or ethnicity (U.S.Census Bureau 2003). A combined approach also allows SES data to be appended to persons assigned Hispanic or Asian race/ethnicity.

Last, health plans have begun to exploit the logic of combined geocoding and surname analyses to examine potential disparities in care. Several plans, participating in AHRQ/RWJF's National Health Plan Learning Collaborative to Reduce Disparities and Improve Quality, are using geocoding

Table 3: Advantages and Limitations of Geocoding and Surname Analysis for Assigning Race and Ethnicity

Geocoding	*Surname Analysis*
Advantages	
Can be implemented quickly and inexpensively	Can be implemented quickly and inexpensively
Nonintrusive	Nonintrusive
Unaffected by response bias	Unaffected by response bias
Useful for imputing race data individual-level data such as self-report are unavailable	Useful for estimating language or religious affiliation for certain groups (particularly when combined with geocoding)
Useful for oversampling African Americans for surveys	Useful for oversampling Hispanic and Asian subgroups groups for surveys
Useful for estimating proportion of health plan membership who are African American	Useful for estimating proportion of health plan membership who are African American
Can be used to identify racial/ethnic disparities in care	Can be used to identify racial/ethnic disparities in care
Can provide an estimate of SES	Can be used to impute ethnicity when data are missing from self-report data
Can be used to estimate geographical access to care	
Can be used to inform geographic based outreach or education, or community interventions	
Limitations	
Accuracy depends on degree of racial segregation in the area	Accuracy depends on the ethnic group's concentration among residents in the area
Less accurate for Hispanics and Asians	Less accurate for married women and adopted persons
Inapplicable to Native American populations outside of reservations	Less accurate for members of certain subpopulations, e.g., Cubans, Puerto Ricans, Filipinos, and Hawaiians. Uknown accuracy for Native Americans
Accuracy depends on geographic scale of geocoding (census tract, block group, block)	Inapplicable to African Americans
Accuracy decreases when used to determine racial/ethnic characteristics of a specific individual or small group	Ethnicity may be unassigned for a small portion of names
Race may be unassignable for 10 percent or more of enrollees	Lists used to identify Asian subpopulations cannot be used in aggregate to identify all Asians without adjusting for the differing sensitivities of the lists for each subpopulation
Accuracy can vary across vendors (or software for geocoding)	Accuracy can vary across vendors (or software for geocoding)
May not be well-suited for quality improvement interventions targeted at individuals as well as populations (e.g., disease management programs)	May not be well-suited for quality improvement interventions targeted at individuals as well as populations (e.g., disease management programs)
Requires some programming expertise and access to Census data if vendor or commercial software are not used	Requires some expertise in matching names to list, more so, if first and middle names are used
Geocoded measures may capture independent effects associated with the individual as well as their neighborhood	

and surname analyses to estimate enrollee race/ethnicity as an initial approach to examine and address disparities (Agency for Healthcare Research and Quality 2004). Preliminary analyses suggest that although sensitivity of these indirect measures of race/ethnicity is low in some service areas, such as the West Coast, the positive and negative predictive values are generally high for blacks, Hispanics, Asians, and nonblacks across different service areas. In other words, although geocoding/surname analyses may significantly underidentify minority enrollees in some plans, there is a reasonably high degree of certainty that those enrollees assigned to one of the four racial/ethnic groups based on geocoding/surname analyses actually belong to that racial/ethnic group. Initial results also suggest that indirect measures enable plans to identify patterns of disparities for different racial/ethnic groups on key quality indicators (e.g., HEDIS diabetes measures) that are consistent with analyses based on more direct measures of race/ethnicity. An added benefit of these approaches is that it is a relatively easy step to use the geocoded information to create geospatial maps that highlight areas of large disparities and their characteristics, which in turn can help plans make decisions about how to most effectively target their efforts to reduce disparities. Much more work is planned as part of the Collaborative's activities to confirm and clarify these preliminary findings and to better delineate appropriate use and interpretation of disparities analyses based on indirect measures of race/ethnicity.

GETTING STARTED

Table 4 lists the basics steps involved with geocoding and surname analyses. Health care organizations interested in using these indirect methods can either perform their coding in-house or contract out with commercial vendors. Most vendors set pricing based on the number of addresses or names and the number of variables that the organization wishes to have appended (e.g., race, ethnicity, SES, language, religion, and so on). Some vendors charge based on the number of matches. Typical fees for basic processing of 250,000 addresses or names or addresses range from as low as $2.00 per 1,000 to as high as $15.00 with additional fees for others variables (e.g., SES). Most vendors offer volume discounts; some offer discounts for repeat customers.

Coding accuracy varies considerably between vendors and does not correlate with costs. Organizations are encouraged to submit a test file to vendors and review match rates (>90 percent addresses matched to Census Block Group is good), repeatability and accuracy before signing a contact (Whitsel

Table 4: Basic Steps in Conducting Geocoding and Surname Analysis

Geocoding	*Surname Analysis*
1. Decide on health plan measures for which race/ethnicity, SES, neighborhood characteristics are needed	1. Decide on health plan measures for which race/ethnicity, SES, neighborhood characteristics are needed
2. Decide on racial/ethnic groups to be distinguished (see Census Bureau website for guidance)	2. Decide on ethnic groups to be distinguished
3. Select a feasible level of geocoding (Census tract, block-group, or block) and sample (entire plan versus selected communities)	3. Select a feasible level of analysis (Census tract, block-group, or block) and sample (entire plan versus selected communities)
4. Weigh "in-house" analysis versus commercial vendor alternatives. For vendors: submit test file for validation and compare results	4. Decide whether to use existing name lists or software ("in-house" analysis) or use commercial vendor. When using vendors, shop around, submit test file for validation
5. Format address file and/or use Zip+4 then submit to vendor to assign Census code (e.g., block-group)	5. Use Census data to estimate racial and ethnic composition of area
6. Select desired race/ethnicity or SES variables from Census data files	6. Format name file and submit to vendor or use program to match last names with list and assign probability for different ethnicities
7. Set cut-points, compute measures to be used	7. Establish cut-point depending on purpose for which data will be used
8. Link racial/ethnic and SES data derived from geocoding to individuals and performance data	8. Link ethnicity data to individuals and performance data

et al. 2004). Similarly, because the accuracy of these methods varies by segregation levels and concentration of minorities in particular census tract, block-groups, or blocks, the overall accuracy of these methods will likely vary between health care organizations and between different regions of the country. For this reason, it is advisable for plans that have some self-reported race/ethnicity data to compare estimates of racial/ethnic composition and disparities in care based on self-report data with those based on indirect race/ethnicity data. Such validation allows health care organizations to determine whether accuracy across regions is sufficient to assess disparities using indirect data.

CONCLUSION

Combined geocoding and surname analysis offers health plans a timely means to infer race/ethnicity among their plan members for the purpose of assessing

disparities in health care processes and outcomes. Although self-report represents the gold standard, indirect methods (suitably validated for a sample of plan members) offer a defensible interim alternative in lieu of direct data.

A combined approach can yield positive predictive and negative predictive values of roughly 80 and 90 percent, respectively, thereby offering a viable means for assigning race and ethnicity for purposes of examining disparities in care until self-reported data can be systematically collected on all plan members.

ACKNOWLEDGMENTS

We appreciate the insightful comments of Peter Morrison, Ph.D., regarding earlier versions of this manuscript. This work was supported by the Robert Wood Johnson Foundation. Neither author has any conflict of interest.

REFERENCES

Abrahamse, A., P. A. Morrison, and N. M. Bolton. 1994. "Surname Analysis for Estimating Local Concentration of Hispanic and Asians." *Population Research and Policy Review* 13: 383–98.

Agency for Healthcare Research and Quality. 2004. "Major Health Plans and Organizations Join AHRQ to Reduce Racial and Ethnic Disparities in Health Care" [accessed March 5, 2005]. Available at http and 2005 www.ahrq.gov/news/press/pr2004/dispcolpr.htm

Andjelkovich, D. A., R. B. Richardson, P. E. Enterline, and R. J. Levine. 1990. "Assigning Race to Occupational Cohorts Using Census Block Statistics." *American Journal of Epidemiology* 131: 928–34.

Bickart, B., and D. Schmittlein. 1999. "The Distribution of Survey Contact and Participation in the United States: Constructing a Survey-Based Estimate." *Journal of Marketing Research* 36: 286–94.

Blustein, J. 1994. "The Reliability of Racial Classifications in Hospital Discharge Abstract Data." *American Journal of Public Health* 84: 1018–21.

Chen, J. T., N. Krieger, S. K. Van Den Eeden, and C. P. Quesenberry. 2002. "Different Slopes for Different Folks: Socioeconomic and Racial/Ethnic Disparities in Asthma and Hay Fever among 173,859 U.S. Men and Women." *Environmental Health Perspectives* 110 (suppl): 211–6.

Chen, W., D. B. Petitti, and S. Enger. 2004. "Limitations and Potential Uses of Census-Based Data on Ethnicity in a Diverse Community." *Annals of Epidemiology* 14: 339–45.

Choi, B. C., A. J. Hanley, E. J. Holowaty, and D. Dale. 1993. "Use of Surnames to Identify Individuals of Chinese Ancestry." *American Journal of Epidemiology* 138: 723–34.

Coldman, A. J., T. Braun, and R. P. Gallagher. 1988. "The Classification of Ethnic Status Using Name Information." *Journal of Epidemiology and Community Health* 42: 390–5.

Coronado, G. D., T. D. Koepsell, B. Thompson, S. M. Schwartz, R. S. Wharton, and J. E. Grossman. 2002. "Assessing Cervical Cancer Risk in Hispanics." *Cancer Epidemiology, Biomarkers and Prevention* 11: 979–84.

Cummins, C., H. Winter, K. K. Cheng, R. Maric, P. Silcocks, and C. Varghese. 1999. "An Assessment of the Nam Pehchan Computer Program for the Identification of Names of South Asian Ethnic Origin." *Journal of Public Health Medicine* 21: 401–6.

Elliott, M., A. Fremont, N. Lurie, P. A. Morrison, P. Pantajo, and A. Abrahamse. 2005. "A New Method for Estimating Racial/Ethnic Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity." Unpublished Analyses.

Elo, I. T., C. M. Turra, B. Kestenbaum, and B. R. Ferguson. 2004. "Mortality among Elderly Hispanics in the United States: Past Evidence and New Results." *Demography* 41: 109–28.

Elreedy, S., N. Krieger, P. B. Ryan, D. Sparrow, S. T. Weiss, and H. Hu. 1999. "Relations between Individual and Neighborhood-Based Measures of Socioeconomic Position and Bone Lead Concentrations among Community-Exposed Men: The Normative Aging Study." *American Journal of Epidemiology* 150: 129–41.

Fiscella, K., and P. Franks. 2001. "Impact of Patient SES on Physician Profiles: A Comparison of Census Derived and Individual Measures." *Medical Care* 39: 8–14.

Fremont, A. M., A. S. Bierman, S. L. Wickstrom, C. E. Bird, M. M. Shah, J. J. Escarce, and T. S. Rector. 2005. "Use of Indirect Measures of Race/Ethnicity and Socioeconomic Status in Managed Care Settings to Identify Disparities in Cardiovascular and Diabetes Care Quality." *Health Affairs* 24: 516–26.

Fremont, A., and N. Lurie. 2004. *The Role of Race and Ethnic Data Collection in Eliminating Disparities in Health Care.* Washington, DC: National Academy Press.

Fremont, A., P. Pantajo, M. Elliott, P. A. Morrison, A. Abrahamse, and N. Lurie. 2005. *Use of Indirect Measures of Race/Ethnicity to Examine Disparities in Managed Care.* AcademyHealth Annual Research Conference, Chicago.

Fremont, A., and S. L. Wickstrom. 2002. "Socioeconomic, Racial/Ethnic, and Gender Differences in Quality and Outcomes of Care as It Relates to Cardiovascular Disease." Final Report. Rockville, MD: AHRQ.

Fremont, A., S. L. Wickstrom, and J. J. Escarce. 2003. "Does Differential Diffusion of Innovations Contribute to Disparities in Health Care?" Final Report. Rockville, MD: AHRQ.

Geronimus, A. T., and J. Bound. 1998. "Use of Census-Based Aggregate Variables to Proxy for Socioeconomic Group: Evidence from National Samples." *American Journal of Epidemiology* 148: 475–86.

Glaeser, E. L., and J. L. Vigdor. 2001. *Racial Segregation in the Census 2000: Promising News.* Washington, DC: Brookings Institute Center on Urban & Metropolitan Policy.

Hage, B. H., R. G. Oliver, J. W. Powles, and M. L. Wahlqvist. 1990. "Telephone Directory Listings of Presumptive Chinese Surnames: An Appropriate Sampling Frame for a Dispersed Population with Characteristic Surnames." *Epidemiology* 1: 405–8.

Harding, S., H. Dews, and S. L. Simpson. 1999. "The Potential to Identify South Asians Using a Computerised Algorithm to Classify Names." *Population Trends* 97: 46–9.

Harland, J. O., M. White, R. S. Bhopal, S. Raybould, N. C. Unwin, and K. G. Alberti. 1997. "Identifying Chinese Populations in the UK for Epidemiological Research: Experience of A Name Analysis of the FHSA Register. Family Health Services Authority." *Public Health* 111: 331–7.

Hassett, P. 2005. "Taking on Racial and Ethnic Disparities in Health Care: The Experience at Aetna." *Health Affairs* 24: 417–20.

Heck, K. E., K. C. Schoendorf, and G. F. Chavez. 2002. "The Influence of Proximity of Prenatal Services on Small-for-Gestational-Age Birth." *Journal of Community Health* 27: 15–31.

Institute of Medicine. 2002. *Unequal treatment: Confronting Racial and Ethnic Disparities in Health Care*, edited by B. D. Smedley, A. Y. Stith, and A. R. Nelson. Washington, DC: National Academy Press.

Jacobs, E. A., and D. S. Lauderdale. 2001. "Receipt of Cancer Screening Procedures Among Hispanic and Non-Hispanic Health Maintenance Organization Members." *Cancer* 91 (suppl): 257–61.

Kressin, N. R., B. H. Chang, A. Hendricks, and L. E. Kazis. 2003. "Agreement between Administrative Data and Patients' Self-Reports of Race/Ethnicity." *American Journal of Public Health* 93: 1734–9.

Krieger, N. 1992. "Overcoming the Absence of Socioeconomic Data in Medical Records: Validation and Application of a Census-Based Methodology." *American Journal of Public Health* 82: 703–10.

Krieger, N., J. T. Chen, P. D. Waterman, M. J. Soobader, S. V. Subramanian, and R. Carson. 2002. "Geocoding and Monitoring of US Socioeconomic Inequalities in Mortality and Cancer Incidence: Does the Choice of Area-Based Measure and Geographic Level Matter?: The Public Health Disparities Geocoding Project." *American Journal of Epidemiology* 156: 471–82.

———. 2003. "Choosing Area Based Socioeconomic Measures to Monitor Social Inequalities in Low Birth Weight and Childhood Lead Poisoning: The Public Health Disparities Geocoding Project (US)." *Journal of Epidemiology and Community Health* 57: 186–99.

Krieger, N., P. D. Waterman, J. T. Chen, M. J. Soobader, and S. V. Subramanian. 2003. "Monitoring Socioeconomic Inequalities in Sexually Transmitted Infections, Tuberculosis, and Violence: Geocoding and Choice of Area-Based Socioeconomic Measures—The Public Health Disparities Geocoding Project (US)." *Public Health Report* 118: 240–60.

Krieger, N., D. R. Williams, and N. E. Moss. 1997. "Measuring Social Class in US Public Health Research: Concepts, Methodologies, and Guidelines." *Annual Review of Public Health* 18: 341–78.

Kwok, R. K., and B. C. Yankaskas. 2001. "The Use of Census Data for Determining Race and Education as SES Indicators: A Validation Study." *Annals of Epidemiology* 11: 171–7.

Lauderdale, D. S., and B. Kestenbaum. 2000. "Asian American Ethnic Identification by Surname." *Population Research and Policy Review* 19: 283–300.

———. 2002. "Mortality Rates of Elderly Asian American Populations Based on Medicare and Social Security Data." *Demography* 39: 529–40.

Lee, R., and D. Sutton. 2002. "Better Ethnic Targeting: New Methodology Enhances Voter Files." *Political Adviser* [accessed on September 27, 2005]. Available at http://www.findarticles.com/p/articles/mi_m2519/is_6_23/ai_8997371

Logan, J. 2001. *Ethnic Diversity Grows, Neighborhood Integration Lags Behind.* Albany, NY: Lewis Mumford Center, University at Albany.

Martineau, A., and M. White. 1998. "What's Not in A Name. The Accuracy of Using Names to Ascribe Religious and Geographical Origin in a British Population." *Journal of Epidemiology and Community Health* 52: 336–7.

Massey, D. S., and N. A. Denton. 1989. "Hypersegregation in U.S. Metropolitan Areas: Black and Hispanic Segregation along Five Dimensions." *Demography* 26: 373–91.

NAACCR Expert Panel on Hispanic Identification. 2003. *Report of the Expert Panel on Hispanic Identification.* Springfield, IL: North American Association of Central Cancer Registries.

Nanchahal, K., P. Mangtani, M. Alston, and I. Santos Silva dos. 2001. "Development and Validation of a Computerized South Asian Names and Group Recognition Algorithm (SANGRA) for Use in British Health-Related Studies." *Journal of Public Health Medicine* 23: 278–85.

National Quality Forum. 2002. *Improving Healthcare Quality for Minority Patients.* Washington, DC: National Quality Forum.

Nerenz, D. R. 2005. "Health Care Organizations' Use of Race/Ethnicity Data to Address Quality Disparities." *Health Affairs* 24: 409–16.

Nerenz, D. R., V. L. Bonham, R. Green-Weir, C. Joseph, and M. Gunter. 2002. "Eliminating Racial/Ethnic Disparities in Health Care: Can Health Plans Generate Reports?" *Health Affairs* 21: 259–63.

Nicoll, A., K. Bassett, and S. J. Ulijaszek. 1986. "What's in a Name? Accuracy of Using Surnames and Forenames in Ascribing Asian Ethnic Identity in English Populations." *Journal of Epidemiology and Community Health* 40: 364–8.

Perez-Stable, E. J., R. A. Hiatt, F. Sabogal, and R. Otero-Sabogal. 1995. "Use of Spanish Surnames to Identify Latinos: Comparison to Self-Identification." *Journal of the National Cancer Institute Monographs* 18: 11–5.

Perkins, R. C. 1993. *Evaluating the Passel-Word Spanish Surname List: 1990 Decennial Census Post Enumeration Survey Results.*, Population Division Working Paper No. 4. Washington, DC: Population Division, U.S. Bureau of the Census.

Perot, R. T., and M. Youdelman. 2001. *Racial, Ethnic, and Primary Language Data Collection in the Health Care System: An Assessment of Federal Policies and Practices.* New York: The Commonwealth Fund.

Robinson, L., and A. Grant-Thomas. 2004. *Race, Place, and Home: A Civil Rights and Metropolitan Opportunity Agenda.* Cambridge, MA: Civil Rights Project, Harvard University.

Rosenbaum, S., and M. W. Painter. 2005. *Assessing the Legal Implications of Using Health Data to Improve Health Care Quality and Eliminate Health Care Disparities.* Washington, DC: George Washington University/Robert Wood Johnson Foundation.

Rosenwaike, I. 1994. "Surname Analysis as Means of Estimating Minority Elderly." *Research on Aging* 16: 212–27.

Rosenwaike, I., and B. S. Bradshaw. 1988. "The Status of Death Statistics for the Hispanic Population of the Southwest." *Social Science Quarterly* 69: 722–36.

Rosenwaike, I., K. Hempstead, and R. G. Rogers. 1991. "Using Surname Data in U.S. Puerto Rican Mortality Analysis." *Demography* 28: 175–80.

Sheth, T., M. Nargundkar, K. Chagani, S. Anand, C. Nair, and S. Yusuf. 1997. "Classifying Ethnicity Utilizing the Canadian Mortality Data Base." *Ethnicity and Health* 2: 287–95.

Swallen, K. C., S. L. Glaser, S. L. Stewart, D. W. West, C. N. Jenkins, and S. J. McPhee. 1998. "Accuracy of Racial Classification of Vietnamese Patients in a Population-Based Cancer Registry." *Ethnicity and Disease* 8: 218–27.

Swallen, K. C., D. W. West, S. L. Stewart, S. L. Glaser, and P. L. Horn-Ross. 1997. "Predictors of Misclassification of Hispanic Ethnicity in a Population-Based Cancer Registry." *Annals of Epidemiology* 7: 200–6.

U.S. Census Bureau. 2003. *Census, Table 1. Language Use, English Ability, and Linguistic Isolation for the Population 5 Years and over by State: 2000.* Washington, DC: U.S. Census Bureau.

———. 2004. *Health Insurance Coverage 2003.* Washington, DC: U.S. Census Bureau.

Weech-Maldonado, R., L. S. Morales, M. Elliott, K. Spritzer, G. Marshall, and R. D. Hays. 2003. "Race/Ethnicity, Language, and Patients' Assessments of Care in Medicaid Managed Care." *Health Services Research* 38: 789–808.

Whitsel, E. A., K. M. Rose, J. L. Wood, A. C. Henley, D. Liao, and G. Heiss. 2004. "Accuracy and Repeatability of Commercial Geocoding." *American Journal of Epidemiology* 160: 1023–9.

Winkleby, M. A., and B. Rockhill. 1992. "Comparability of Self-Reported Hispanic Ethnicity and Spanish Surname Coding." *Hispanic Journal of Behavioral Sciences* 14: 487–95.

Word, D. L., and R. C. Perkins. 1996. "Building a Spanish Surname List for the 1990's—A New Approach to an Old Problem." Technical Working Paper No. 13. Washington, DC: Population Division, U.S. Bureau of the Census.

Measuring Health Care Quality. 2004. *Obstacles and Opportunities.* Washington, DC: Workgroup on Quality: National Committee on Vital and Health Statistics.

Zaslavsky, A. M., L. B. Zaborski, and P. D. Cleary. 2002. "Factors Affecting Response Rates to the Consumer Assessment of Health Plans Study Survey." *Medical Care* 40: 485–99.